Universidade Cruzeiro do Sul

NOVA DIDÁTTICA
EDUCAÇÃO E DESENVOLVIMENTO

Inferência Estatística III

Análises de Variâncias

Prof. Dr. Juliano van Melis

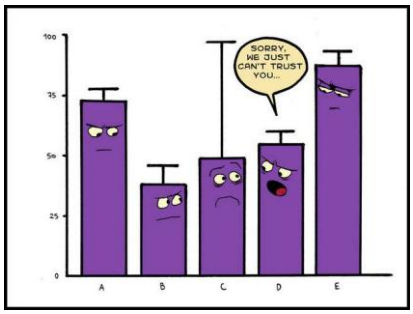
Parte II

Parte 1 — Conteúdo

- **Análise da Variância – ANOVA**
  - Introdução
  - Pressupostos da ANOVA
  - Teste de Levine para homogeneidade da variância
  - Estatística F para testar igualdade de várias médias
  - Interpretação do Quadro ANOVA
    - Outputs R, SPSS e SAS
  - ANOVA de um fator com o MS Excel® e R
  - ANOVA com dois fatores com MS Excel® e R
  - ANOVA com medidas repetidas
- **Teste de Kruskall-Wallis**

Parte 2 — Conteúdo

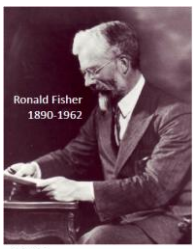
- **Correlação Linear Simples**
  - Coeficiente Correlação Linear de Pearson
  - Significância da correlação linear
  - Medida de associação paramétrica
  - Teste t student para análise da significância CLP
  - Aplicações e análises com MS Excel® e R
- **Medida de associação não-paramétrica**
  - Teste de Spearman
- **Correlação Bisserial**
- **Avaliação**



Análises de Variâncias

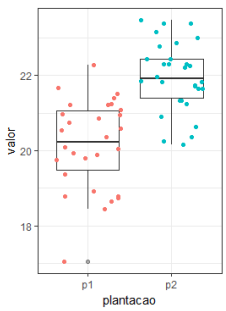
RELEMBRANDO ANOVA

Definição: "Técnica estatística usada para determinar se as amostras de dois ou mais grupos surgem de populações com médias iguais. A análise de variância é empregada para uma medida dependente"



ANOVA
Maximum likelihood estimation

Variável Dependente (métrica)	~	Variável Independente (categórica)



$$x_{ij} = \overline{x_{\cdot}} + erro_{ij} + grupo_{\cdot j}$$

erro "DENTRO" do grupo *j*

erro "ENTRE" os grupos

$$x_{ij} = \bar{x}_{.} + \underbrace{(\bar{x}_i - \bar{x}_{.})}_{\text{erro "ENTRE" os grupos}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\text{erro "DENTRO" do grupo } j}$$

Soma dos Desvios Quadrados ENTRE os grupos
"Sum of Squares ***BETWEEN***"

$$\mathrm{SSD}_B = \sum_i \sum_j (\bar{x}_i - \bar{x}_{.})^2 = \sum_i n_i (\bar{x}_i - \bar{x}_{.})^2$$

Soma dos Desvios Quadrados DENTRO dos grupos
"Sum of Squares ***WITHIN***"

$$\mathrm{SSD}_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

É possível normalizar a soma dos quadrados, calculando a ***média dos desvios quadrados***

$$\mathrm{MS}_W = \mathrm{SSD}_W / (N - k)$$
$$\mathrm{MS}_B = \mathrm{SSD}_B / (k - 1)$$

Mean Squares *Within* & *Between*

N: número total
k: número de grupos

Pressupostos

1. **Resíduos** (*erros*) devem seguir uma **distribuição normal**: $\boldsymbol{erro_{ij} \sim N(0, \sigma^2)}$
2. **Homogeneidade das variâncias**: As contribuições das variâncias dos grupos devem ser equivalentes para a variância total.
3. **Amostras independentes**: a observação de uma variável não pode influenciar outra observação. Atenção para medidas repetidas!

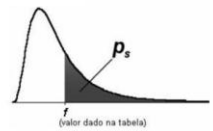
$$F = \mathrm{MS}_B / \mathrm{MS}_W$$

Se o valor de **F = 1**
→ Médias dos Quadrados ***ENTRE*** os grupos é semelhante às Médias dos Quadrados ***DENTRO*** dos grupos.
Se o valor de **F < 1**
→ Médias dos Quadrados ***ENTRE*** os grupos é menor que as Médias dos Quadrados ***DENTRO*** dos grupos.
→ Nesses dois casos, as variâncias dentro dos grupos é tão grande que sobressaem a qualquer sinal que os grupos tenham.

Se o valor de F > 1
→ Médias dos Quadrados ***ENTRE*** os grupos é maior que as Médias dos Quadrados ***DENTRO*** dos grupos.
→Nesse caso, **os grupos parecem ter papel importante para a variação dos valores.**

$$F = \mathrm{MS}_B / \mathrm{MS}_W$$

DISTRIBUIÇÃO F



	p	Degrees of freedom in numerator (df1)										
		1	2	3	4	5	6	7	8	12	24	1000
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	60.71	62.00	63.30
	0.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.2
	0.025	647.8	799.5	864.2	899.5	921.8	937.1	948.2	956.6	976.7	997.3	1017.8
	0.010	4052	4999	5404	5624	5764	5859	5928	5981	6107	6234	6353
	0.001	405312	499725	540257	562868	576496	586033	593195	597954	610362	623703	636101
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.41	9.45	9.49
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.41	19.45	19.49
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.41	39.46	39.50
	0.010	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.42	99.46	99.50
	0.001	998.38	998.84	999.31	999.31	999.31	999.31	999.31	999.31	999.31	999.31	999.31

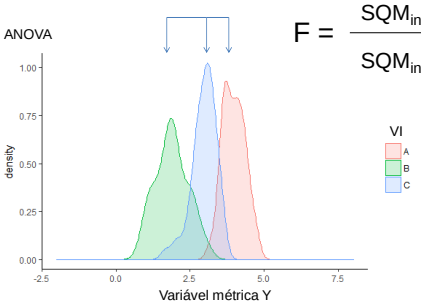
EXEMPLO

```
> anova(mod)
Analysis of Variance Table

Response: valor
           Df Sum Sq Mean Sq F value    Pr(>F)
plantacao  1 61.453  61.453  52.029 1.285e-09 ***
Residuals 58 68.506   1.181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```


TABELA ANOVA 1-FATOR

Fonte da Variação	SQ	gl	Variância	Razão F
Entre	SQE	k - 1	$S^2_{entre} = \frac{SQE}{k-1}$	$F = \frac{S^2_{entre}}{S^2_{dentro}}$
Dentro	SQD	n - k	$S^2_{dentro} = \frac{SQD}{n-k}$	
Total	SQT = SQE+SQD	n - 1		



$$F = \frac{SQM_{intergrupos}}{SQM_{intragrupos}}$$


```
> anova(aov(y~A))
Analysis of Variance Table

Response: Variável métrica Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
VI          2 117.466  58.733  204.17 < 2.2e-16 ***
Residuals 147  42.288   0.288
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

H_0: Médias dos grupos são iguais

E depois?

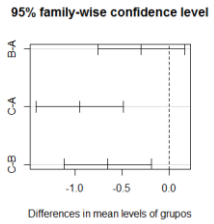
→ Fazer Tukey HSD (ou outro teste *post hoc*)

Teste Tukey HSD (Honest Significant Difference)

$$HSD = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{MS_{within}}{n_{grupo}}}}$$

→ É realizado ***APÓS*** ANOVA

As diferenças são significativas quando os "intervalos" não estão encostados no eixo 0.

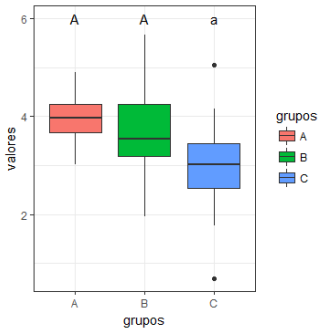



```
> TukeyHSD(anova_mod)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = valores ~ grupos, data = dados_1)

$grupos
            diff        lwr        upr     p adj
B-A -0.2981098 -0.7608306  0.1646110 0.2791641
C-A -0.9520277 -1.4147486 -0.4893069 0.0000127
C-B -0.6539179 -1.1166388 -0.1911971 0.0031928

> plot(TukeyHSD(anova_mod))
>
```



Escolhe-se um padrão de **letras diferentes** (ou tamanho de letras diferentes) para "agrupar" médias semelhantes.

→ Neste caso, grupos **A** e **B** são semelhantes, enquanto que o grupo **C** é diferente dos grupos **A** e **B**

ETAPAS NECESSÁRIAS PARA EFETUAR UMA ANOVA

1. Verifique se os dados contínuos seguem uma distribuição normal

```
shapiro.test(y)
```

2. Verifique o pressuposto de homocedasticidade

```
var.test(y ~ x)
bartlett.test(y ~ x)
levene.test(y ~ x)
```

3. Variáveis são independentes?

+ Número amostral semelhante
+ Amostragem suficiente

TESTAR VALIDADE DA ANOVA

1. ANOVA significativa, execute um teste *post hoc*

```
TukeyHSD(aov(y ~ x))
```

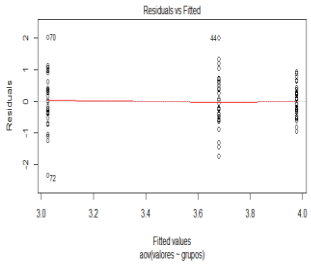
2. Cheque as homogeneidade das variâncias

```
plot(aov(y ~ x), which=1)
```

3. Distribuição Normal dos resíduos

```
plot(aov(y ~ x), which=2)
```

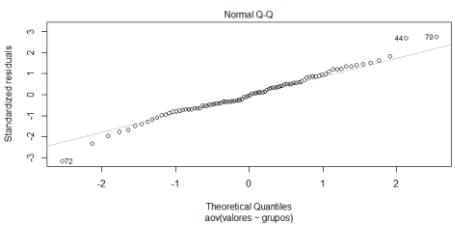
→ Considere Teste Não-paramétrico: **Kruskal-Wallis**



Residuals (*within*)

- Devem apresentar valores médios próximo de 0 (Linha vermelha na horizontal, no 0)
- Variâncias semelhantes e homogêneas (distribuição dos pontos no eixo y deve ser parecida)
- Eixo x mostra as médias dos 3 grupos (valores ajustados – *fitted*)

→ **Resíduos com distribuição normal**
→ **Números destacados são possíveis *outliers*.**



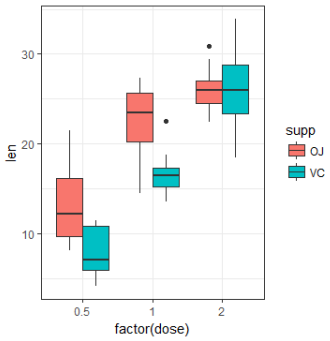
→ Quando há o interesse de verificar a relação entre ***duas variáveis categóricas*** em relação a uma ***variável contínua***

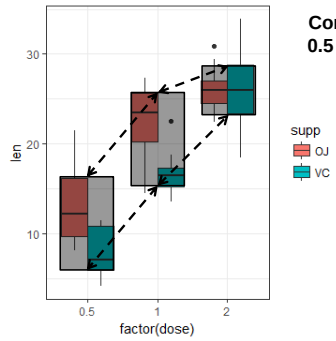
$$y_1 \sim x_1 + x_2$$

```
data("ToothGrowth")
?ToothGrowth
dente <- ToothGrowth
```

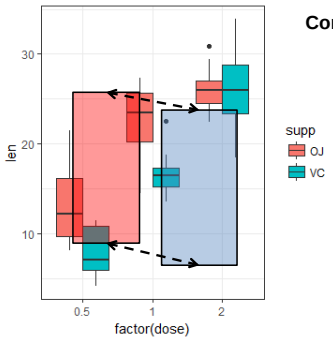
len ~ supp + dose

len [numérico]: Comprimento do Dente ("Tooth length")
supp [fator]: Tipo de suplemente ("Supplement type"):
VC: Vitamine C
OJ: Orange Juice.
dose [numeric]: Dose em mg/dia ("Dose in milligrams/day")





Comparando as doses:
0.5 *versus* **1.0** *versus* **2.0**



Comparando as fontes:
OJ *versus* **VC**

supp \ dose	0.5	1.0	2.0
VC	4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7	16.5 16.5 15.2 17.3 22.5 17.3 13.6 14.5 18.8 15.5	23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5
OJ	15.2 21.5 17.6 9.7 14.5 10 8.2 9.4 16.5 9.7	19.7 23.3 23.6 26.4 20 25.2 25.8 21.2 14.5 27.3	25.5 26.4 22.4 24.5 24.8 30.9 26.4 27.3 29.4 23

QUADRO ANOVA 2-FATORES

Fonte da variação	Soma dos Quadrados	Graus de liberdade	Quadrados Médios QM (variâncias)	F
Linhas	SSD_{linhas}	Linhas-1	$SSD_{linhas}/Linhas\text{-}1$	QM_{lin}/QM_{res}
Colunas	$SSD_{colunas}$	colunas-1	$SSD_{colunas}/col\text{-}1$	QM_{col}/QM_{res}
Linhas:Colunas	SSD_{inter}	(l-1)(c-1)	$SSD_{inter}/(l\text{-}1)(c\text{-}1)$	QM_{int}/QM_{res}
Resíduos	$SSD_{resíduos}$	l.c.(n'-1)	$SSD_{inter}/l.c.(n'\text{-}1)$	
TOTAL	SSD_{total}	n-1		

Soma dos Desvios Quadrados

$$SSD_{colunas} = m\sum_j \left(\overline{x_{\bullet j}} - \overline{x_{\bullet\bullet}}\right)^2 \qquad SSD_{linhas} = n\sum_i \left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)^2$$

$$SSD_{\text{int}} = \sum_i \sum_j \left(\overline{x_{ij}} - \overline{x_{i\bullet}} - \overline{x_{\bullet j}} + \overline{x_{\bullet\bullet}}\right)^2 \qquad SSD_{TOTAL} = \sum \left(x_{ij} - \overline{x_{\bullet\bullet}}\right)^2$$

$\overline{x_{\bullet\bullet}}$: Média geral

$\overline{x_{ij}}$: Média da linha ***i*** na coluna ***j***

$\overline{x_{\bullet j}}$: Média da coluna ***j***

$\overline{x_{i\bullet}}$: Média da linha ***i***

m : número de elementos para cada coluna ***j***

n : número de elementos para cada linha ***i***

```
> ANOVA_int <- aov(len ~ factor(dose) + supp + factor(dose):supp, data= dente)
> summary(ANOVA_int)
                Df Sum Sq Mean Sq F value   Pr(>F)
factor(dose)     2 2426.4  1213.2  92.000  < 2e-16 ***
supp             1  205.4   205.4  15.572 0.000231 ***
factor(dose):supp  2  108.3    54.2   4.107 0.021860 *
Residuals       54  712.1    13.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ANOVA_int <- aov(len ~ factor(dose)*supp, data= dente)
> summary(ANOVA_int)
                Df Sum Sq Mean Sq F value   Pr(>F)
factor(dose)     2 2426.4  1213.2  92.000  < 2e-16 ***
supp             1  205.4   205.4  15.572 0.000231 ***
factor(dose):supp  2  108.3    54.2   4.107 0.021860 *
Residuals       54  712.1    13.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

→Com interação entre fatores
→Quando a interação **não é significativa**, levar em consideração somente **modelo aditivo**

INTERPRETAÇÃO

- A dose recebida (dose) influencia nas médias do comprimento dos dentes
  $F_{2,54}$ = 92, p-valor<0.0001
-
- A fonte da vitamina C (supp) recebida influencia nas médias do comprimento dos dentes
  $F_{1,54}$ = 15.57, p-valor<0.0001

- A relação entre dose e comprimento dos dentes é influenciada pela fonte de vitamina C
  $F_{2,54}$ = 4.11, p-valor<0.05

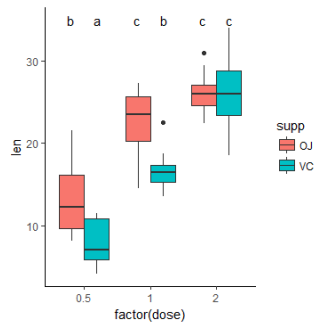

```
> TukeyHSD(ANOVA_int)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = len ~ factor(dose) * supp, data = dente)

$`factor(dose)`
      diff       lwr       upr   p adj
1-0.5  9.130  6.362488 11.897512 0.0e+00
2-0.5 15.495 12.727488 18.262512 0.0e+00
2-1    6.365  3.597488  9.132512 2.7e-06

$supp
      diff       lwr       upr     p adj
VC-OJ -3.7 -5.579828 -1.820172 0.0002312

$`factor(dose):supp`
              diff       lwr        upr     p adj
1:OJ-0.5:OJ    9.47   4.671876 14.2681238 0.0000046
2:OJ-0.5:OJ   12.83   8.031876 17.6281238 0.0000000
0.5:VC-0.5:OJ -5.25 -10.048124 -0.4518762 0.0242521
1:VC-0.5:OJ    3.54  -1.258124  8.3381238 0.2640208
2:VC-0.5:OJ   12.91   8.111876 17.7081238 0.0000000
2:OJ-1:OJ      3.36  -1.438124  8.1581238 0.3187361
0.5:VC-1:OJ  -14.72 -19.518124 -9.9218762 0.0000000
1:VC-1:OJ     -5.93 -10.728124 -1.1318762 0.0073930
2:VC-1:OJ      3.44  -1.358124  8.2381238 0.2936430
0.5:VC-2:OJ  -18.08 -22.878124 -13.2818762 0.0000000
1:VC-2:OJ     -9.29 -14.088124 -4.4918762 0.0000069
2:VC-2:OJ      0.08  -4.718124  4.8781238 1.0000000
1:VC-0.5:VC    8.79   3.991876 13.5881238 0.0000210
2:VC-0.5:VC   18.16  13.361876 22.9581238 0.0000000
2:VC-1:VC      9.37   4.571876 14.1681238 0.0000058
```

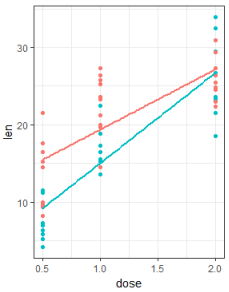
ANCOVA
ANálise de **COVA**riância

```
> ANCOVA_ <- aov(len ~ dose*supp, data= dente)
> summary(ANCOVA_)
             Df Sum Sq Mean Sq F value   Pr(>F)
dose          1 2224.3  2224.3 133.415  < 2e-16 ***
supp          1  205.3   205.3  12.317 0.000894 ***
dose:supp     1   88.9    88.9   5.333 0.024631 *
Residuals    56  933.6    16.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Y ~ X1*cov(X2)

Numérico Fator Numérico

ANCOVA
ANálise de **COVA**riância



Diferença entre os ***b*** dos ***Fatores***?
len ~ a + ***b****dose [supp=OJ]
len ~ a + ***b****dose [supp=VC]

```
> if(!require(lsmeans)){install.packages("lsmeans")}
> lstrends(ANCOVA_, 'supp', var='dose')
 supp dose.trend       SE df lower.CL upper.CL
 OJ     7.811429 1.195422 56 5.416711 10.20615
 VC    11.715714 1.195422 56 9.320997 14.11043

Confidence level used: 0.95
>
```

+ DISCIPLINA 11

Prós
- Custa menos (precisa de menos sujeitos)
- Maior poder estatístico

Contra:
- Princípio da independência
- Precisa estar em ordem (“Tempo1”, ”Tempo2” ...)
- Valores faltantes

Dados “WIDE” ⇔ Dados “LONG”

ID	Fator	Temp1	Temp2
1	Trata	0.1	0.2
2	Control	1.1	1.2
...			
n	FatorX	5.1	5.2

ID	Fator	Tempo	Valor
1	Trata	1	0.1
1	Trata	2	0.2
2	Control	1	1.1
2	Control	2	1.2
...			
n	FatorX	2	5.2

+ DISCIPLINA 11

Kruskal-Wallis

→É alternativa não-paramétrica para o teste ANOVA para um fator (one-way ANOVA).
→Semelhante ao Teste U de Wilcoxon pois também utiliza ***ranking***.

Características
- Análise de variância não paramétrica
- 3 ou + grupos independentes
- Hipótese: As distribuições de todos os grupos são iguais,
- Hipótese: As medianas de todos os grupos são iguais
- Insensível a outliers
- Os grupos não precisam ter o mesmo tamanho

Kruskal-Wallis

ALTERNATIVA NÃO-PARAMÉTRICA PARA ANOVA *ONE WAY*

Cuidados
- Se a distribuição for normal é melhor usar o teste ANOVA de um critério (one-way)
- Precisa ter 4 ou mais elementos na amostra de cada grupo
- Se tiver só 2 grupos use o Mann-Whitney

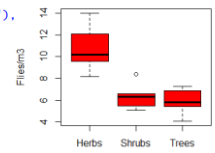
Kruskal-Wallis

```
> Herbs <- c(14, 12.1, 9.6, 8.2, 10.2)
> Shrubs <-c(8.4, 5.1, 5.5, 6.6, 6.3)
> Trees <- c(6.9, 7.3, 5.8, 4.1, 5.4)
> flies <- data.frame(local=c(rep('Herbs',5),
+                             rep('Shrubs',5),
+                             rep('Trees',5)),
+                     obs=c(Herbs, Shrubs, Trees))
> kruskal.test(obs~local, flies)

        Kruskal-Wallis rank sum test

data:  obs by local
Kruskal-Wallis chi-squared = 8.72, df = 2, p-value = 0.01278

> boxplot(Herbs, Shrubs, Trees,
+         names = c('Herbs', 'Shrubs', 'Trees'),
+         ylab = "Flies/m3", col = 'red')
>
```



Friedman

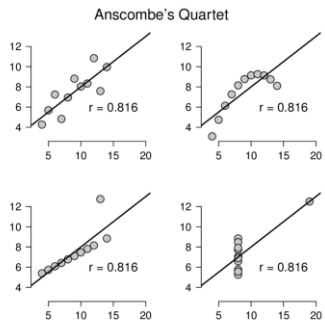
ALTERNATIVA NÃO-PARAMÉTRICA PARA ANOVA *TWO WAY*

→Alternativa para ANOVA com dois fatores (two-way ANOVA)
→Equivalente ao ***Teste de Sinais***, onde testa pares de + ou – dentro de cada par.
→É menos sensível que o teste de sinais de Wilcoxon

```
> sample_n(CO2,5)
Grouped Data: uptake ~ conc | Plant
   Plant        Type  Treatment conc uptake
77   Mc2 Mississippi    chilled 1000   14.4
69   Mc1 Mississippi    chilled  675   22.2
37   Qc3      Quebec    chilled  175   21.0
63   Mn3 Mississippi nonchilled 1000   27.8
42   Qc3      Quebec    chilled 1000   41.4
> friedman.test(uptake~conc|Plant, data = CO2)

        Friedman rank sum test

data:  uptake and conc and Plant
Friedman chi-squared = 59.677, df = 6, p-value = 5.236e-11
```



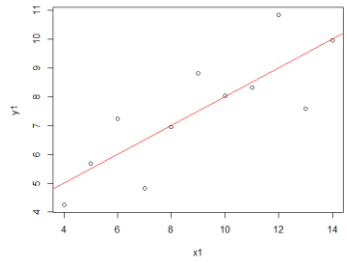
REGRESSÃO E CORRELAÇÃO LINEAR SIMPLES

Suponha que você queira descrever a relação entre duas ***variáveis contínuas***: y e x, onde temos observações independentes individuais i.
Assumiremos que a relação entre essas duas variáveis é uma correlação linear, logo, assumimos a equação simples a seguir:

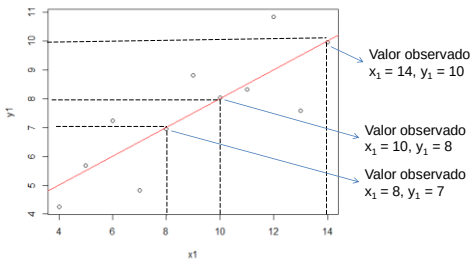
$$y_i = \alpha + \beta x_i + \epsilon_i$$

Onde:
y : variável dependente
x: variável dependente
α: intercepto da equação linear
β: inclinação da equação linear (***coeficiente da regressão***)
ε: erro {~ N(0, σ^2)}

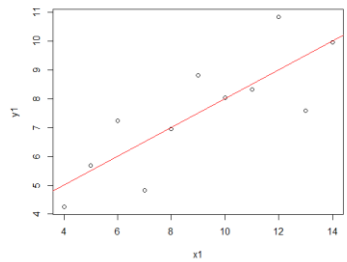

```
data("anscombe")
plot(y1~x1, anscombe)
abline(lm(y1~x1, anscombe), col="red")
```




```
data("anscombe")
plot(y1~x1, anscombe)
abline(lm(y1~x1, anscombe), col="red")
```

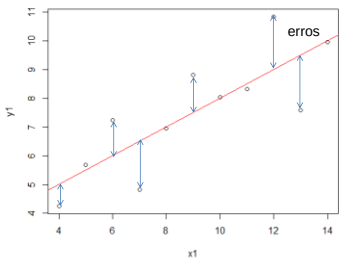


Conforme aumenta x, aumenta y

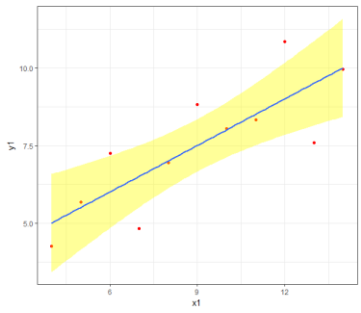


x	y
8	3 + 0.5*8
10	3 + 0.5*10
14	3 + 0.5*14

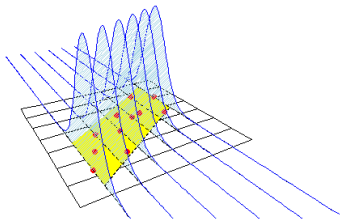
y = 3 + 0.5*x + erro



y = 3 + 0.5*x + erro



erro ~ N (0, σ^2)



Como estimar α, β e σ^2?

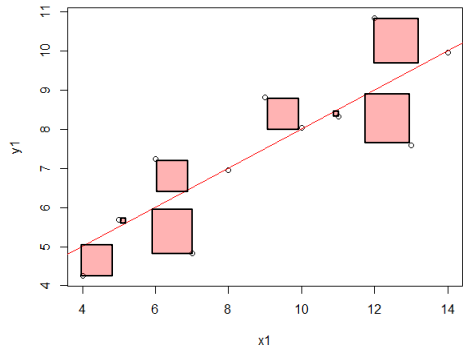
→Método dos Mínimos Quadrados

(*Ordinary Least Squares* ou *Method of Least Squares*)

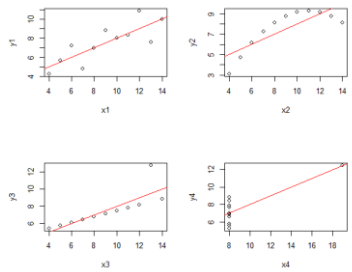
Encontrar valores de alfa e beta
que minimizem a
soma dos quadrados dos resíduos (SS)

$$SS_{res} = \sum_i (y_i - (\alpha + \beta x_i))^2$$

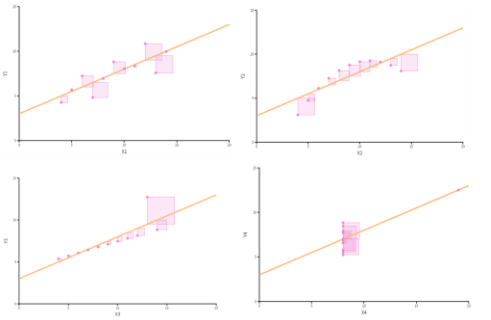
→Método dos Mínimos Quadrados



→Método dos Mínimos Quadrados



→Método dos Mínimos Quadrados



http://students.brown.edu/seeing-theory/regression-analysis/index.html#section1

→Método dos Mínimos Quadrados **EXERCÍCIO**

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Microsoft Office Excel

1. Abrir: “anscombe.xls”
2. Calcular Inclinação da reta (β)
3. Calcular intercepto (α)
4. Calcular Soma dos Desvios Quadrados

→Método dos Mínimos Quadrados

	n	$\bar{x}$	$\bar{y}$	$\hat{B}_0$	$\hat{B}_1$	SSE
Model	11	9.00	7.50	3.00	0.50	13.76

http://students.brown.edu/seeing-theory/regression-analysis/index.html#section1

Pressupostos

-As amostras são representativas da população para que predição seja feita
-O erro é uma variável aleatória com média zero
-As variáveis independntes são medidas sem erros
-Os preditores são linearmente independentes (é possível expressar qualquer preditor como uma combinação linear dos outros)
-Os erros **não apresentam correlação** (a matriz de variância-covariância dos erros é diagonal e cada elemento diferente de zero é a variância do erro)
-A variância do erro é constante ao longo das observações (**homocedasticidade**)

Consequências de um Modelo Errado/Ruim

-Atinge a teoria por trás
-Coeficientes ineficientes ou enviesados (colinearidade, baixa significância) levam a interpretações errôneas (apesar de sua teoria)
-*Outliers* implicam que você não está apto para utilizar o seu modelo para toda a sua população
-*Overfitting = overconfidence*

Como saber se meu modelo é Ruim/Errado?

-Coeficiente de determinação
-Gráfico dos resíduos
-Teste de White (ou de Breusch-Pagan)
-VIF (Variance Inflation Factor)
-Leverage Points
-Cook`s Distance
-Teste de Outliers
Ver: https://www.statmethods.net/stats/rdiagnostics.html
Próximas disciplinas: Regressão Linear, Regressão Múltipla..

Hipótese Nula: Inclinação é igual a zero:

β = 0

$$t = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}$$

graus de liberdade = n - 2

Hipótese Nula: Inclinação é igual a zero:

β = 0

$$t = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}$$

$$\text{s.e}(\hat{\beta}) = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2 \,/\, (\text{n} - 2)}{\sum(\bar{x} - x_i)^2}}$$

Hipótese Nula: Inclinação é igual a zero:

β = 0

$$t = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}$$

RSS: Residuals Sum of Squares

$$\text{s.e}(\hat{\beta}) = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{(n-2)} \Big/ \sum(\bar{x} - x_i)^2}$$

Graus de Liberdade (também de t)

SS_x: Soma dos Desvios Quadrados de x


```
> summary(lm(y1~x1, anscombe))

Call:
lm(formula = y1 ~ x1, data = anscombe)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,    Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

>
```


Correlação (r: amostral)

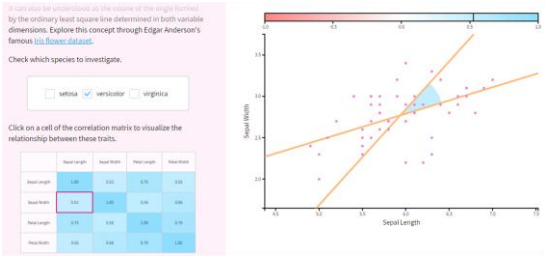
É uma medida de relação linear entre duas variáveis. É definida pela fórmula a seguir e apresenta valores entre +1 e -1:

$$r = \frac{SS_{xy}}{\sqrt{SS_x} \cdot \sqrt{SS_y}}$$

Correlação (ρ: populacional)

Pode ser entendida como o valor do coseno d ângulo formado pela linha entre ambas as retas das duas dimensões:

y ~ x e x ~ y



Correlação (ρ e *r*)

Teste de Hipótese – Significância do coeficiente de correlação

Hipótese nula: ρ = 0

Hipótese alternativa: ρ ≠ 0

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

graus de liberdade = n - 2

Correlação (ρ e *r*)

Termo associado com a **covariância**

Covariância: significa co-variação, ou seja, como duas variáveis variam de forma conjunta

$$r = \frac{SS_{xy}}{\sqrt{SS_x} \cdot \sqrt{SS_y}}$$

Correlação (ρ e *r*)

Termo associado com a **covariância**

Covariância: significa co-variação, ou seja, como duas variáveis variam de forma conjunta

$$r = \frac{SS_{xy}/n}{\sqrt{SS_x/n} \cdot \sqrt{SS_y/n}}$$

Definição e Introdução | Coeficiente de Pearson e Significância | Medida de associação paramétrica | Medida de associação não-paramétrica

Correlação (ρ e *r*)

Termo associado com a **covariância**

Covariância: significa co-variação, ou seja, como duas variáveis variam de forma conjunta

$$r = \frac{SS_{xy}/n}{\sqrt{SS_x/n} \cdot \sqrt{SS_y/n}}$$

$\mathbf{cov_{xy}}$

$\mathbf{var_x}$ $\mathbf{var_y}$

Definição e Introdução | Coeficiente de Pearson e Significância | Medida de associação paramétrica | Medida de associação não-paramétrica

Matriz de Variância-Covariância

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Z) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) & \mathrm{Cov}(Y,Z) \\ \mathrm{Cov}(X,Z) & \mathrm{Cov}(Y,Z) & \mathrm{Var}(Z) \end{bmatrix}$$

→Muitas aplicações estatísticas calculam a matriz de variância-covariância para os estimadores de parâmetros em um modelo estatístico.
→ Importante para ***análises multivariadas***

Definição e Introdução | Coeficiente de Pearson e Significância | Medida de associação paramétrica | Medida de associação não-paramétrica

Medidas de Associação

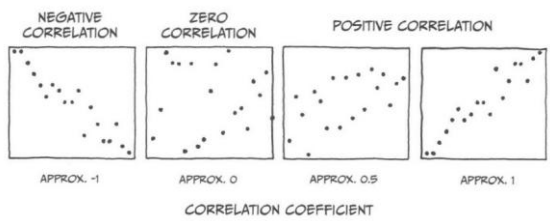
Data types	Index	Value range	Formula
Numerical and numerical	Correlation coefficient	-1 - 1	$\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \times \sum(y-\bar{y})^2}} = \frac{Sxy}{\sqrt{Sxx \times Syy}}$
Numerical and categorical	Correlation ratio*	0 - 1	$\frac{\text{interclass variance}}{\text{intraclass variance + interclass variance}}$
Categorical and categorical	Cramer's coefficient*	0 - 1	$\sqrt{\frac{\chi_0^2}{\text{the total number of values} \times (\min[\text{the number of lines in the cross tabulation, the number of rows in the cross tabulation}] - 1)}}$

* See page 121, "Correlation Ratio," and page 127, "Cramer's Coefficient."

Fonte: Takahashi. 2008.The Manga Guide to Statistics

Definição e Introdução | Coeficiente de Pearson e Significância | Medida de associação paramétrica | Medida de associação não-paramétrica

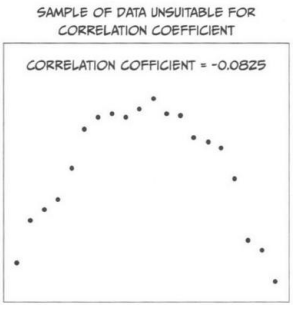
Medidas de Associação



Fonte: Takahashi. 2008.The Manga Guide to Statistics

Definição e Introdução | Coeficiente de Pearson e Significância | Medida de associação paramétrica | Medida de associação não-paramétrica

Medidas de Associação



Fonte: Takahashi. 2008.The Manga Guide to Statistics

Correlação - Observações importantes:

• O fato de duas variáveis serem fortemente correlacionadas não implica uma relação direta de ***causa e efeito*** entre elas.
→Não se pode afirmar que X causa Y e nem que Y causa X.

• É possível que a correlação entre as duas variáveis possa ser causada por uma ***terceira variável***, ou combinação de diversas outras variáveis. (exemplo: número de bares com número de igrejas em uma cidade)

• É possível que uma correlação forte entre duas variáveis seja apenas coincidência (***relações espúrias***).

Coeficiente de Determinação (R^2)

•Medida de ajuste a uma regressão linear

•É a porcentagem explicada pelo modelo linear

•R^2 = Variação explicada/Total da variação

•É sempre entre 0 e 100%

•0% indica que o modelo não explica a variabilidade dos dados dependentes em relação a média

•100% indica que o modelo explica toda a variabilidade dos dados dependentes em torno da média.

Coeficiente de Determinação

•Em algumas áreas é comum que R^2 seja baixo. Por exemplo, em psicologia (R^2 de 50% é alto)

•Se seu ajuste (R^2) for baixo, mas você tem preditores significativo, conclusões podem ser feitas acerca das mudanças dos valores dos preditores estarem associados a mudanças nos valores da resposta

•Quando há muitas variáveis, utiliza-se o R^2 ajustado.

•N: número amostral, n número de variáveis independentes e m o número de observações necessárias para conseguir uma boa precisão do modelo → $N = m^n$

Coeficiente de Determinação - Limitações

•R^2 não consegue determinar quando as estimativas ou predições são enviesadas, por isso analisar os resíduos (gráficos) é muito importante

•R^2 não indica se o modelo da regressão é adequado. Você pode ter um R^2 baixo mas um bom modelo ou ter um R^2 alto e um modelo que não está adequado aos dados

•O R^2 estimado por você é o R^2 da população, com viés.

Coeficiente de Determinação

O que o R^2 indica:

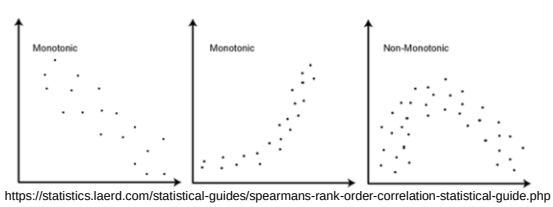
1. Proporção de variabilidade total explicada pelo modelo;
2. Melhoria quanto ao modelo nulo
3. É o quadrado da correlação (correlação varia de -1 a 1, coeficiente de determinação de 0 a 1).

Correlação (ρ ou rho) de Spearman (r_s)

→Usa um coeficiente de correlação em ranking

O coeficiente de correlação de postos de Spearman (r_s) é uma medida da força da relação entre duas variáveis:
• Pode ser utilizado para descrever relações lineares e não lineares entre dados;
• Não requer que as populações de cada variável sejam normalmente distribuídas.
• O problema é que a interpretação não é tão clara.
→ **Utilizado para descrever relações monotônicas**

Correlação (ρ ou rho) de Spearman (r_s)



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Correlação de Spearman (r_s)

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

QUANDO TEM EMPATES

EXERCÍCIO

Microsoft Office Excel

1. Abrir "notas_spearman.xls"
2. Calcular r_s
3. Testar a hipótese de que a correlação é igual a zero entre as duas discplinas
4. Concluir

Correlação de Spearman (r_s) EXERCÍCIO

1. Use `data("USArrests")`
2. Primeiro veja a relação e avalie a correlação entre População urbana (`UrbanPop`) e Estupro (`Rape`)
   a. Graficamente: `plot(UrbanPop ~ Rape, USArrests)`
   b. Utilizando a função , veja a correlação de Spearman entre População urbana (`UrbanPop`) e Estupro (`Rape`) : `cor.test(x = USArrests$UrbanPop, y = USArrests$Rape, method = "spearman")`
3. Agora faça o mesmo para a relação entre população urbana (`UrbanPop`) e Assassinatos (`Murder`)
4. Conclua

Correlação Bisserial (r_b ou r_{pb})

→Usado para relação entre uma variável dicotômica (1 ou 0, por ex) com uma variável contínua (naturalmente ou forçadamente)
→Também varia de -1 a +1

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{pq} \qquad s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

M_1 = Média do grupo "1".
M_0 = Média do grupo "0"
s_n = Desvio padrão para todo o teste.
p = Proporção de casos no grupo "0".
q = Proporção de casos no grupo "1".

! Para amostras usa-se *n-1* !

Correlação Bisserial (r_b ou r_{pb})

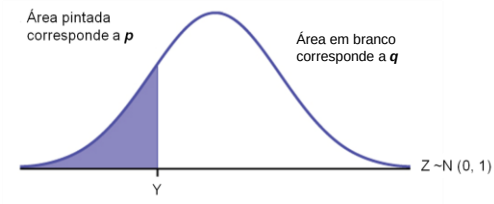
→Quando os grupos são artificiais (variável contínua que foi transformada em binomial, utiliza-se o r_b. Caso contrário, utiliza-se a correlação ponto-bisserial
→**É** um coeficiente de Pearson "modificado"

$$r_b = \frac{(Y_1 - Y_0)\cdot\left(\frac{pq}{Y}\right)}{\sigma_y}$$

Y_0 : Média dos valores para valores de x=0,
Y_1 : Média dos valores para valores de x=1,
q : Proporção de dados para x=0,
p : Proporção de dados para x=1,
σ_y : Desvio padrão da população.

Correlação Bisserial (r_b ou r_{pb})

$$r_b = \frac{(Y_1 - Y_0)\cdot\left(\frac{pq}{Y}\right)}{\sigma_y}$$



Correlação Bisserial (r_b ou r_{pb})

EXERCÍCIO

1. Use `data("mtcars")`
2. Depois utilize a função `cor.test()` nas variáveis `am` (variável dicotômica) e `mpg` (variável contínua)
3. Conclua

→Não é normalmente utilizada.
→**Problemas**: precisa apresentar os mesmos pressupostos de uma relação linear (homocedasticidade, independência), além de ser pouco robusta.
→Portanto, é mais utilizado ANOVA ou teste t para comparar valores médios
"Exceção": Mostra a força de uma associação.

AVALIAÇÃO INDIVIDUAL

Definição e Introdução | Coefiente de Pearson e Significância | Medida de associação paramétrica | Medida de associação não-paramétrica

Correlação Bisserial (r_b ou r_{pb})

→TESTE DE HIPÓTESE
H0: Não há correlação entre grupos e valores observados

$$t_0 = r\sqrt{\frac{n-2}{1-r^2}}$$

EXERCÍCIO

1. Abrir "sexo.xls"
2. Calcular r_{pb}
3. Testar a hipótese de que a correlação é igual a zero entre "sexo" e "sexo"
4. Concluir